DOCUMENT RESUME

ED 259 559 FL 015 119

AUTHOR Stemberger, Joseph Paul
TITLE Overview of Adult Speech Error Corpora.
PUB DATE Apr 85
NOTE 9p.; "Transcript Analysis" is the newsletter of the Child Language Data Exchange System at Carnegie-Mellon University.
PUB TYPE Information Analyses (070) -- Journal Articles (080)
JOURNAL CIT Transcript Analysis; v2 n1 p17-24 Apr 1985

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; *Databases; Data Collection; Directories; Dutch; English; *Error Patterns; German; Information Storage; Japanese; Language Research; Languages; Norwegian; *Oral Language; Portuguese; *Research Projects; *Structural Analysis (Linguistics); Swedish; Thai

ABSTRACT

A listing of substantial, known adult speech error corpora includes seven major corpora in English, 11 in languages other than English (German, Swedish, Dutch, Norwegian, Portuguese, Japanese, and Thai), information on seven collectors of smaller English samples, and two references for anecdotal samples, one in French and one in Welsh. Each major listing includes the collector's name and address, the language, corpus size, error types included, collection methods, transcription status, and availability to the public. Notes on the definition of standard collection procedures are also included. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Transcript Analysis
Dept. of Psychology
Carnegie-Mellon University

Volume 2 no 1 1985
April 1985

ED259559

4. Overview of Adult Speech Error Corpora

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
X Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Joseph Paul Stemberger

Department of Psychology

Indiana University

Bloomington, Indiana

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Brian MacWhinney
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Researchers in child language have long been aware that adult native speakers of a language make many errors in spontaneous speech that resemble the errors made by children in many ways. Interest in systematically comparing the errors made by children and adults seems to be increasing, with the goal of determining in what ways the errors are alike and in what ways they differ. The following list of corpora is being provided to facilitate research in this area. As a general rule of thumb, the larger the corpus is, the more uses it can be put to. I have not made an effort to locate and list very small corpora (under 1000), but will list a few of the more interesting or readily available corpora of that size.

For those interested in pursuing the published literature on adult speech errors, Anne Cutler has a bibliography that lists most papers available as of early 1982:

Cutler, Anne. (1982) Speech errors: a classified bibliography. Bloomington, IN: Indiana University Linguistics Club. (34 pages)

The following list will be organized by language, and will list the size of the corpus, the names of the people who collected/are collecting it, the method of collection, the transcription status, and the current availability of the data to the general public. The corpora include all types of errors unless otherwise noted. Since most of the corpora were collected in a very similar fashion, let me begin by defining a standard collection procedure.

STANDARD COLLECTION PROCEDURES: The collector monitors speech carefully, trying to detect all errors that occur. This is usually concurrent with normal language processing, and the collector is usually trying to process the speech for meaning as well, whether taking part in a conversation or listening to a lecture. Only the errors of adult native speakers of the language are collected. Care is taken to ensure that the errors did not occur while the speaker was reading aloud. On detecting an error, the collector writes it down immediately, with as much of the context as possible, the name of the speaker (if known), and the date.

FL015119

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
X Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

BEST COPY AVAILABLE

ENGLISH CORPORA

1. The UCLA Corpus

Collector: Victoria A. Fromkin
Department of Linguistics
UCLA
Los Angeles, CA. 90024
Language: English (mostly American)
Size: 8000+ errors
Error Types: all types
Collection Method: Standard collection procedures. Additionally, speakers are asked for information about what they intended to say and the causes of the error, if possible. Errors are accepted from other listeners, but are kept separate.
Transcription Status: Originally handwritten on cards, at least a portion of the corpus is now on the computer.
Availability To Public: Unknown.

2. The MIT Corpus

Collectors:

Merrill Garrett	Stefanie Shattuck-Hufnagel
Department of Psychology	Research Laboratory of Electronics
MIT	MIT
Cambridge, MA 02139	Cambridge, MA 02139

Language: English (mostly American)
Size: 6000+ errors
Error Types: all types
Collection Methods: Standard collection procedures.
Transcription Status: Until recently, handwritten on cards. Currently being typed into computer textfiles.
Availability To Public: Currently not, but copies will be available once the corpus has been computerized.

3. The Fay-Cutler Corpus

Collectors:

David Fay	Anne Cutler
Bell Laboratories	MRC Applied Psychology Unit
6A-304A	15 Chaucer Road
Naperville-Wheaton Road	Cambridge, CB2 2EF
Naperville, IL 60566	ENGLAND

Language: English (mostly American)
Size: 5000+ errors
Error Types: all types
Collection Methods: Standard collection procedures. Cutler also has a tape-recorded corpus of several hundred errors in addition.
Transcription Status: Unknown. Probably hand-written or typed.
Availability To Public: Unknown.

4. **The Stemberger Corpus**

Collector: Joe Stemberger
Department of Psychology
Indiana University
Bloomington, Indiana
Language: English (mostly American)
Size: 7200+ errors
Error Types: all types
Collection Methods: Standard collection procedures. To minimize intrusiveness, speakers are never asked about the error. The collector developed a "speech error mode of perception", and was either actively searching for errors or was not. The decision to collect was always made in advance, and no errors were collected without that decision, no matter how interesting they appeared to be. No errors were collected unless enough attention could be devoted to the task to ensure accuracy. These measures were included as a weak control for attention biases. No errors are accepted from other listeners.
Transcription Status: All handwritten.
Availability to Public: Currently not, though you can ask and would probably be allowed access if you come in person.

5. **The Toronto Corpus**

Collector: Peter Reich
Department of Psychology
University of Toronto
Toronto, Canada M5S 1A1
Language: English (Canadian)
Size: 4000 errors
Error Types: all types
Collection Methods: Errors were collected by 200 students in psycholinguistics courses at the University of Toronto. All were familiar with linguistic categories, but none had received any explicit training in speech errors. As a class project, they were required to collect errors for one month. Each student used standard collection procedures.
Transcription Status: Unknown, but probably handwritten or typed
Availability To Public: Unknown.

6. **The Harley Corpus**

Collector: Trevor A. Harley
Department of Psychology
University of Dundee
Dundee DD1 4HN
SCOTLAND
Language: English (British & Scottish)
Size: 1100+ errors
Error Types: all types
Collection Procedures: Standard collection procedures. Speakers who did not correct their errors were asked about them, as in Fromkin's corpus.

Transcription Status: Unknown. but probably handwritten or typed.
Availability To Public: Unknown.

7. **The London-Lund Corpus**

Collectors: Alan Garnham. Richard Shilcock. Gordon Brown.
Andrew Mill. Anne Cutler
Laboratory of Experimental Psychology
University of Sussex
Falmer
Brighton BN1 9QG
ENGLAND
Language: English (British)
Size: 191 errors
Error Types: all types
Collection Methods: 34 passages of 5000 words each from spontaneous speech where the speaker was unaware of being tape recorded were transcribed onto the computer. Transcriptions were carefully examined for errors. with the collectors deciding as a group whether there was any way to interpret each instance as a correct utterance. Effort was made to detect all errors. but no claim for exhaustiveness is made. since the tapes on which the transcriptions were based were sometimes not clear enough to determine with certainty what the speaker said.
Transcription Status: see on availability
Availability To Public: All detected errors appear. with analysis. in the article Garnham. A.. Shillcock. R.C.. Brown. G.D.A.. Mill. A.I.D.. & Cutler. A. (1981) Slips of the tongue in the London-Lund corpus of spontaneous conversation. Linguistics 19. 805-817 (Also reprinted in Cutler. A. (1982) Slips of the tongue and language production. The Hague: Mouton.)

8. **Smaller Corpora.**

Several smaller corpora of errors in English have been collected and reported on. all cited in Cutler's bibliography. The collectors are:

Jean Aitchison & Miron Straf
H.H. Bawden
Donald S. Boomer & John D.M. Laver
Mymon Goldstein
W.H.N. Hotopf
E.W. Roberts
Ryan D. Tweney. Sharon Tkacz. & Sally Zabura

LANGUAGES OTHER THAN ENGLISH

1. **The Meringer Corpus**

Collector: Rudolf Meringer
Language: German (Viennese)
Size: ca. 2800 errors (plus child, aphasic & other errors)
Error Types: all types
Collection Methods: Standard collection procedures. Asked speakers about the errors if not self-corrected.
Transcription Status: see below
Availability To Public: A large part of Meringer's corpus is published in two books, though it is not clear how large a part:

Meringer, R., & Mayer, K. (1895) Versprechen und Verlesen. Stuttgart: Goschensche Verlag. (Reissued in 1978, Amsterdam: John Benjamins.

Meringer, R. (1908) Aus dem Leben der Sprache. Berlin: Behr's Verlag.

The first book is printed using the Gothic style alphabet.

2. **The Berg Corpus**

Collector: Thomas Berg
Lindenstr. 1
6721 Freisbach
West Germany
Language: German
Size: 4300+ errors
Error Types: all types
Collection Methods: Standard collection procedures.
Transcription Status: Unknown, probably handwritten or typed.
Availability To Public: Currently not available

3. **The Bierwisch Corpus**

Collector: Manfred Bierwisch
Language: German
Size: Unknown.
Error Types: all types
Collection Methods: Unknown, presumably standard collection procedures.
Transcription Status: Unknown.
Availability To Public: Unknown.

4. **The Soderpalm Corpus**

Collector: Ewa Soderpalm
Department of Phonetics

University of Lund
Kavlingavagen 20
S-222 40 Lund
SWEDEN

Language: Swedish (Southern)
Size: 209 errors
Error Types: all types
Collection Methods: Standard collection procedures. One fourth of her errors were reported to her by other listeners.
Transcription Status: see availability
Availability To Public: Published in book form:
Soderpalm, E. (1979) Speech errors in normal and pathological speech. (Travaux de l'Institut de Linguistique de Lund. #14.) Malmo: Gleerup.

5. **The Linell Corpus**

Collector: Per Linell
Department of Communication Studies
Linkoping University
S-581 83 Linkoping
SWEDEN

Language: Swedish
Size: 1000+ errors
Error Types: all types
Collection Methods: Standard collection procedures.
Transcription Status: Handwritten.
Availability To Public: Currently not.

6. **The Nooteboom Corpus**

Collector: Sieb Nooteboom
Department of Linguistics
Leyden University
Leyden
THE NETHERLANDS

Language: Dutch
Size: 600+ errors
Error Types: ? Perhaps just phonological errors.
Collection Methods: Presumably standard collection procedures.
Transcription Status: Unknown.
Availability to Public: Unknown.

7. **The Foldvik Corpus**

Collector: Arne Kjell Foldvik
Department of Linguistics
University of Trondheim
N-7055 Dragvoll

NORWAY

Language: Norwegian
Size: 122 errors
Error Types: all types
Collection Methods: Standard collection procedures.
Transcription Status: Unknown.
Availability To Public: Unknown.

8. **The Johns-Lewis Corpus**

Collector: Catherine Johns-Lewis
Department of Modern Languages
University of Aston
Gosta Green
Birmingham B4 7ET
ENGLAND

Language: Portuguese
Size: Unknown.
Error Types: Unknown.
Collection Methods: Presumably standard collection procedures.
Transcription Status: Unknown.
Availability To Public: Not available.

9. **The Hiki Corpus**

Collector: Shizuo Hiki
Research Institute of Electrical Communication
Tohuko University
2-1-1
Katahira, Sendai 980
JAPAN

Language: Japanese
Size: 350+ errors
Error Types: all types
Collection Methods: Presumably standard collection procedures.
Transcription Methods: Unknown.
Availability To Public: Unknown.

10. **The Gandour Corpus**

Collector: Jack Gandour
Department of Audiology and Speech Science
Purdue University
West Lafayette, IN 47907

Language: Thai
Size: 350 errors
Error Types: errors on tones
Collection Procedures: Presumably standard collection procedures.
Transcription Status: Unknown.

Availability To Public: Unknown

11. **Anecdotal Samples**

There are several small, more anecdotal collections of errors in languages other than English:

French (Anne Cutler's paper in La Recherche, 1980, 11, 686-692))
Welsh (Meara & Ellis's paper in Linguistics, 1981, 19, 797-804.)

References to minor work and very old work in other languages (including Italian and Arabic) can be found in Cutler's bibliography.